Deep Multimodal Data Fusion

Year: 2024 | Citations: 184 | Authors: Fei Zhao, Chengcui Zhang, Baocheng Geng

---

Abstract

Multimodal Artificial Intelligence (Multimodal AI), in general, involves various types of data (e.g., images, texts, or data collected from different sensors), feature engineering (e.g., extraction, combination/fusion), and decision-making (e.g., majority vote). As architectures become more and more sophisticated, multimodal neural networks can integrate feature extraction, feature fusion, and decision-making processes into one single model. The boundaries between those processes are increasingly blurred. The conventional multimodal data fusion taxonomy (e.g., early/late fusion), based on which the fusion occurs in, is no longer suitable for the modern deep learning era. Therefore, based on the main-stream techniques used, we propose a new fine-grained taxonomy grouping the state-of-the-art (SOTA) models into five classes: Encoder-Decoder methods, Attention Mechanism methods, Graph Neural Network methods, Generative Neural Network methods, and other Constraint-based methods. Most existing surveys on multimodal data fusion are only focused on one specific task with a combination of two specific modalities. Unlike those, this survey covers a broader combination of modalities, including Vision + Language (e.g., videos, texts), Vision + Sensors (e.g., images, LiDAR), and so on, and their corresponding tasks (e.g., video captioning, object detection). Moreover, a comparison among these methods is provided, as well as challenges and future directions in this area.